

Linguistics and Literature Review (LLR)

Volume , Issue 2, October 2015

Journal DOI:

Issue DOI:

ISSN: 2221-6510 (Print) 2409-109X (Online) Journal homepage: <http://journals.umt.edu.pk/llr/Home.aspx>

Urdu Writing Rules for Online Input in PDA's

**Fareeha Anwar
S. Afaq Husain**

To cite to this article: Fareeha Anwar & S. Afaq Husain (2015). Urdu Writing Rules for Online Input in PDA's, *Linguistics and Literature Review* 1(2): 61- 77.

To link to this article:

Published online: October 31, 2015

Article QR Code:



A publication of the
Department of English Language and Literature
School of Social Sciences and Humanities
University of Management and Technology
Lahore, Pakistan

Urdu Writing Rules for Online Input in PDA's

Fareeha Anwar

Department of Computer Science International Islamic University - Islamabad, Pakistan

S. Afaq Husain

Department of Computing Ripah International University - Islamabad, Pakistan

ABSTRACT

For online input, stroke sequence based recognition is generally employed. For this method, the stroke sequence must be uniquely defined for every character/Ligature. Normally, every language has unique writing rules, which are followed by experienced users and recognition engines. However, there are variations in writing style from person to person and place to place. Languages which are written from right to left e.g. Urdu, Arabic, and Persian etc. are complex and have a lot of variations due to fonts and writing style. If rules are not followed properly, the recognition engine is bound to fail. Therefore, proper writing rules are necessary for online recognition of any language based on stroke sequence. There are no published and acknowledged rules available so far for Urdu language. This paper is an effort in accumulating writing rules for 'Nastalique' font for online Urdu recognition engine

Keywords: online Urdu recognition, writing style, Nastalique font and stroke sequence

Introduction

Writing on tablet, PDA or any online input device, generates a sequence of strokes. These strokes are sent to recognition engine which then accepts or rejects the strokes based on predefined rules. Different fonts are available for single language just like roman bases languages. Each font has its own writing rules which vary from each other and also has influence on each other which causes confusion in reception. People writing Nastalique can use Nasakh rules and vice versa is also possible. If stroke sequence is not followed according to predefined writing rules, the rejection rate increases. If user changes the direction of writing stroke; e.g. if a diagonal line starting from top to bottom is written from bottom to top, its shape seems accurate but stroke sequence totally changed so online recognition engine will not be able to recognize it. Therefore to improve the performance of recognition engine, user should follow writing rules. We faced similar problems when developing online recognition engine for Urdu handwritten characters and ligatures (Husain et al., 2007). Most common mistakes were pen up in middle of ligature/ character, writing ligature

starting from opposite sequence; when pen up is required continue writing generating duplicate sequence etc. Therefore need for devising predefined writing rules was felt. Unless these rules are known and the users are trained according to these rules, the recognition engine is bound to fail or give very high failure rate. While searching for writing rules for Urdu language we came across a lack of well published or publicized writing rules. We have devised rules for Urdu language, which will increase recognition rate. These rules will also be guide for new users to learn writing Urdu language or to use computing device for automatic learning just like writing tutorial. There are 38 characters in the Basic Urdu Character set given below

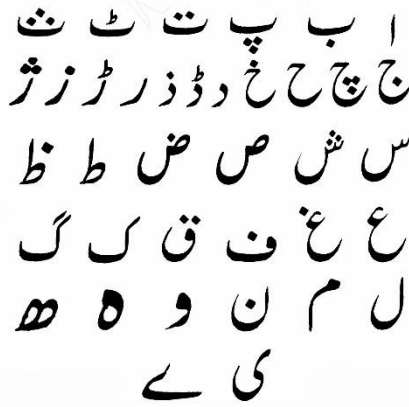


Figure 1. Basic Urdu Alphabets

According to the revised extended character set in Urdu, there are a total 58 Urdu alphabets (Zaheer et al., 2007). The new alphabet set of Urdu is shown in figure 2. Urdu is a cursive languages and very difficult to recognize as discussed in (Starr, 1985).

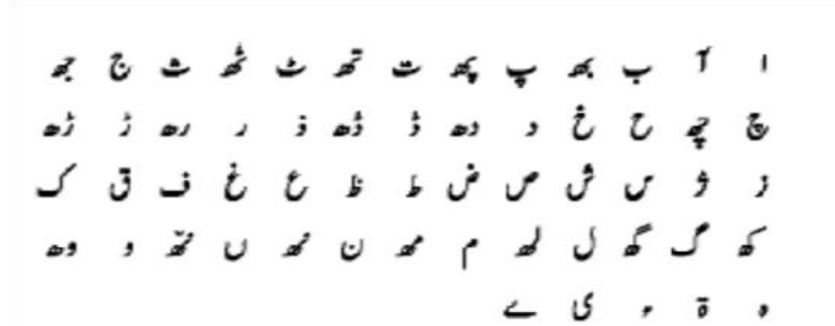


Figure 2. Character Set (58 alphabets) of Urdu Script. (Zaheer et al., 2007)

Four different shapes depending on whether the character is isolated, in the beginning, at the end or connected from both the sides in a word as shown in Table 1 (Reza et al., 2005). Therefore, each character has different shape according to position in a given ligature. Most of the Urdu characters have same shape in ligature provided the same context In the Nastalique way of writing

script, Urdu assumes, so for easy and efficient recognition 38 characters are divided into 18 classes, shown in Table 1.

Table 1. Classification of Urdu

S.N	Urdu Letters	S.N	Urdu Letters		
1	ا	Alif	10	ف ق	Fay
2	ب پ ت ث	Kashti	11	ک گ	Kay
3	ج چ ح خ	Jeem	12	ل	Laam
4	د ڈ ذ	Daal	13	م	Meem
5	ر ژ ز ح	Ray	14	و	Wao
6	س ش	Seen	15	ہ	Gool hay
7	ص ض	Swaad	16	ھ	Do chashmi hay
8	ط ظ	Tuain	17	ی	Choti Yaa
9	ع غ	Aein	18	ے	Bari Yaa

Diacritics

Diacritics are very important in Urdu language. These include diacritics such as *Dots*, *Taaay*, *Hamzaa*, *Diagonal* and *Madaa*, etc.

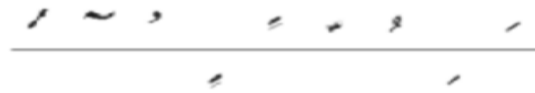


Figure 3. Diacritics/Aerab of Urdu (Aamir et al., 2001)

Basic rules

- Nastalique is actually written from top right to bottom left.

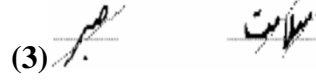


- Each ligature that starts with ک or گ and ends with ی is tilted at approx. 45 degree. This is of particular significance as there is no fixed level or height for any character with respect to base line.

(2)
Chachi
Aunt



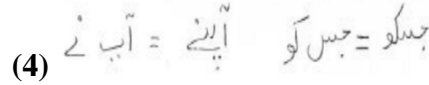
- Some parts of a word are written at an angle of about 30 degrees to the baseline as shown in figure below (Reza et al., 2005).

(3) 

Sabr Salamat

Patience Mercy

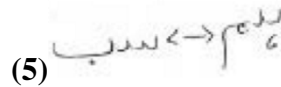
- Write words separately if it is possible as shown in figure below.

(4) 

Aap Nay JisKo

You who ever

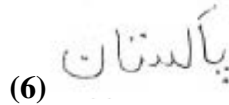
- Writers give distance between two words.

(5) 

Sab Hum

We all

- Long spaces should not be given between ligatures within a single word

(6) 

Pakistan

- All cusps—shosha should be drawn properly. Length of cusp should be proper so that it differentiates the alphabets with in ligature.

(7) 

Shair Liyay

Loin For

- Secondary strokes of first ligature are drawn before second one if word is composed of more than one ligature. 6

5 1
2, 3,

پانی

(8) Pani 4

Water

- Secondary strokes should follow proper sequence. (9)



Shair

Loin

- Stroke written using full qat (length of nib never joins another full qat, rather it always joins with a half *qat* glyph

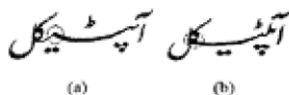


Figure 3 (a). Full qat kashish joined with half qat connector, (b) Full qat kashish joined with full qat circle

Shapes of classes

In each class, there is one or more than one character. Each differs in number and shape of secondary strokes.

Alif class

- Shape of Alif is long vertical line (8 to 10 pixel); direction is from top to bottom. ا
- Shape alif remains same at the isolated, start, middle, or end position of any ligature.
- It remains isolated when comes at start of any ligature.
- Stroke direction changes when it comes after other character i.e., bottom to top instead from top to bottom. ک
- When Laam is followed by alif, ل ا it will be slanting line from top to bottom joining the base of Laam rather than the usual vertical line.

Kashti class

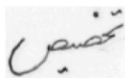
Five characters are present in this class named —Bay ب , —Pay پ , —Taay ت , Ttay ٹ and Saay س.

Basic shape

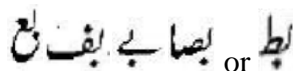
Short vertical line (top to bottom) followed by a ب long horizontal line (right to left), then short vertical line (bottom to top).

Shape at start position

- If jeem class follows ک kashti then its shape is simple middle (5 to 6 pixel) diagonal line at 210 ± 20 degrees (top right to bottom left).



- If kashti is followed by characters having loop in them e.g. —fay, Aien, —qaaf, —wao etc., ray and —yay class,
- Then its shape is simple —ray or half —kashti i.e., short vertical line (top to bottom, 90 ± 20) followed by short horizontal line (right to left, 180 ± 20). There is no cusp after it.



- If meem follows Kashti then its shape will be short curve or diagonal line (top to bottom) at 230 ± 20 degrees



- If —Alif and Kaaf or —Laam follows Kashti then its shape is semicircle (right + downward + left + long upward).
 - For rest it is semicircle (from right + downward + left + upward). There is cusp at the end.



Shape at middle position

If Kashti comes in middle, then it is semicircle (right + downward + left + upward). There is cusp at start as well at end.



All rules for the shape at stating position are same for shape at middle position, including cusp at start.



Shape at end position

At the end position, shape is same as isolated but cusp at start.



Jeem class

Four characters are present in this class named *Jeem* ج, *Chay* چ, *Hay* ح and *Khay* خ.

Basic shape: horizontal line (4 to 5 pixel long) from left to right, followed by downward curve from left to right (angle: downward 210 ± 20 and upward 30 ± 20).

Alternatively, short vertical line from bottom to top (2 to 3 pixel long) at start and rest shape is same as above.



Shape at start position

Half Jeem (before curve) either of two given ways followed by other character ح

Shape at middle position

- Middle diagonal line of 225 ± 20 degrees, followed by sharp edge and diagonal line at 330 ± 20 degrees followed by sharp edge and diagonal line/curve at 210 ± 20 degrees



- If Jeem class follows Kashti, then its shape is half Jeem starts from horizontal line, not from vertical line



Shape at end position

- When it comes at end, its shape is same as isolated, starts from horizontal line, not from vertical line.



Daal class

Three characters are present in this class named —Dal , —Ddal and —Zal.

Basic shape

Medium diagonal at 315 ± 20 degrees from left to right, followed by either downward curve or medium diagonal line at 210 ± 20 degrees from right to left.

Shape at start position

- Remain same as of isolated.

Shape at middle position

- Never comes in middle position of ligature

Shape at end position

- At the end of any ligature it starts with cusp followed by medium vertical line (top to bottom) and then medium horizontal line



Ray class

Four characters are present in this class named, Ray ر, —Aray ا, —Zay ز and ZYaay ي.

Basic shape

Starting from right to left, its shape is medium vertical line of 250 ± 20 degrees, followed by horizontal line of 180 ± 20 degrees

Shape at start position

- Remain same as of isolated.

Shape at middle position

- Never comes in middle position of ligature

Shape at end position

- At the end of any ligature, its shape is medium vertical line at 225 ± 20 (top to bottom) and then medium horizontal line



Seen class

Two characters are present in this class named —Seen س and —Sheen ش

Basic shape

Two small semi circles (right + down+ left + up) having diameter 2 to 3 pixels, with two cusps followed by a big semicircle having diameter 7 to 8 pixels.



Shape at start position

- First part of shape i.e. two semi circles remains same with two cusps There is no last big semicircle. س

- If —Jeem follows —Seen, —Meem or —yaa class then after second semicircle there is no cusp but start the shape of next character.

ج س ی

Shape at middle position

- Start shape remain same but instead of two, three cusps are present including starting cusp as well.
- All conditions of start position remain same.

س س س

Shape at end position

- Remain same as of isolated. Instead of two, it has three cusps (start cusp as well).

س س س

Swad class

Two characters are present in this class named —Swad ص and —Zwad ض

Basic shape

Start from left its shape is diagonal line (2 to 3 pixels) making an angle of 30 ± 10 followed by small downward curve and move back towards starting location making a loop. After starting loop there is small vertical line followed by a cusp and a semicircle (down+ left + up) having diameter 7 to 8 pixels same as last part of —Seen س

Shape at start position

- First part of shape i.e. loop with cusp, remains same. Instead of last semicircle, the shape is small

—Ray with cusp. صیر *Shape*

at middle position

- Whenever it comes in the middle of ligature, there is pen up and shape remains same as of start shape.
- All conditions of start position remain same.

ص ص

(Pen up)

Shape at end position

- Whenever it comes in the end of ligature, there is pen up and shape remains same as of isolated.

(Pen up)

بص
↑

Tua class

Two characters are present in this class named —Tua ط and —Zua ظ

Basic shape

There is straight vertical line from top to bottom making an angle of 270 ± 20 ; same as Alif; followed by a cusp and stroke same as —Dal with curve (not with diagonal line) then move back towards starting location making a loop. After loop, there is small horizontal line ط

Shape at start position

- Shape remains same as of isolated. ط

Shape at middle position

- Whenever it comes in the middle of ligature, there is pen up and shape remains same as of isolated shape. قطر



- All conditions of isolated position remain same.
(Pen up)

Shape at end position

- Whenever it comes in the end of ligature, there is pen up and shape remains same as of isolated. ط
- (Pen up)

Aien class

Two characters are present in this class named —Aienل ع and —Ghaienل غ

Basic shape

Starting from top right making a diagonal line/curve in downward left direction making angle of 225 ± 20 followed by small curve then move back towards right (shape same as —Dall but in opposite direction) and same as small semicircle having diameter 3 to 4 pixels. Then there is a cusp and big semicircle (down-left + right + up) having diameter 7 to 8 pixels

Shape at start position

- First part of shape i.e. small semicircle with cusp remains same. Instead of second curve, there is small horizontal line/curve toward left.

عصا

(Pen up)

- If —Meem or —yaa class follows —Aienل then after first part of shape next character starts at once.

, ,

ع ع م

- If —Aien is followed by —Jeem class then after first part there is small diagonal line at angle 225 ± 20

Shape at middle position

- Whenever it comes in the middle of ligature, there is a small upward diagonal line/curve making an angle 135 ± 20 followed by small horizontal line from left to right and finally small downward diagonal line/curve making an angle 335 ± 20 . This will make a loop.

مجبو

Shape at end position

- At end of ligature, first part remains same as —shape at middle and after loop there is a semicircle(down left + right + up) having diameter 7 to 8 pixels (same as second part of Aien shape without cusp)

بح

Fay class

Two characters are present in this class named —Faay ف and —Qaaf ق

Basic shape

“Faay” Start from down and move towards left, up, right then down make a circle of radius 2 to 3 pixel (loop). After loop, the shape is same as —Kashti. ف

“Qaaf” loop remains same and after loop there is a big semicircle having diameter 7 to 8 pixels (right +down +left+ up). ق

Shape at start position

- If both characters occur at start then the shape is only first part i.e. loop.
- If —Fay is followed by, —Meem م or —yaa ي class then after loop next character starts at once. فم, فم

فم,

- If —Fay is followed by ح then after loop there is small diagonal line at angle 225 ± 20 فح

Shape at middle position

طبقه

- Shape at middle is same as shape at start position only connected with previous character. All conditions remain same as of start position.

Shape at end position

- At the end of ligature the shape is same as of basic shape فف, فطبقه

Kaaf class

Two characters are present in this class named —Kaf ك and —Gaafل گ

Basic shape

At start, there is medium vertical line; instead of short vertical line; rest part is same as basic shape of —Kashti class.

Shape at start position

- If —Jeem, —Ray, —yay class and all the classes having loop follows —Kaaf then its shape is long vertical line (top to bottom, 270 ± 20) followed by short horizontal line (right to left, 180 ± 20). There is no cusp after it كے
- If —Alif or —Laam follows —Kaf class then its shape is same as start position of —Fay class, i.e. loop. All conditions of —Fay at start position remain same.
- For rest it is semicircle (from right + downward + left + upward) with straight vertical line. There is cusp at the end. گ ک

کن

Shape at middle position

- If Kaaf comes in middle, then there is a vertical line from top to bottom making an angle of 90 ± 20 then downward stroke making an angle of 270 ± 20 then small semicircle (right + downward + left + upward). There is cusp at start as well at end. کلب
- If —Alif or —Laam follows —Kaf class then its shape is same as middle position of —Fay class, i.e. loop. All rules for the shape of Fay at middle position remain same

فکل . رکا

Shape at end position

- At the end position, shape is same as isolated but cusp at start.

کلب

Laam class

Two characters are present in this class named —Lam ل and —Nun ن

Basic shape


“Nun”: it is same as last part of —Seen i.e. semi circles (right + down+ left + up) having diameter 7 to 8 pixels.

ن


“Lam”: At start, there is vertical line from top to bottom making an angle of 270 ± 20 , remaining part is same as —Nun.


ل

Shape at start position

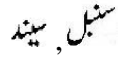
Nun”: At start of any ligature, its shape is same as shape of —Kashti at start position. 


- All conditions associated with shape of —Kashti at start position remain same.

“Lam” at start of any ligature, its shape is same as shape of —Kaaf at start position. 

- All conditions associated with shape of —Kaaf at start position remain same. Except to When —Alif follows —Lam, its shape is vertical line followed by short horizontal line (right to left). 

Shape at middle position

“Nun”: At middle of any ligature, its shape is same as shape of —Kashti at middle position. All conditions associated with shape of —Kashti at middle position remain same. 

“Lam”: At middle of any ligature, its shape is same as shape of —Kaaf at middle position; all conditions associated with shape of —Kaaf at middle position remain same. Except when —Alif follows —Lam, its shape is vertical line followed by short horizontal line (right to left). 

Shape at end position

- At the end position, shape is same as isolated but cusp at start.



Meem class


One character is present named —Meem 

Basic shape

It starts from left. Its shape is diagonal line/curve (2 to 3 pixels) making an angle of 30 ± 10 followed by small downward curve and move back towards starting location making a loop. After starting loop there is small horizontal line followed by a large vertical line from top to bottom same as —Alif.


Shape at start position

Start from up and move towards left, down, right then up make a circle of radius 2 to 3 pixel (loop).


- If —Alif, —Jeem, —Mem or —Yaa class is follows —Mem class then the shape of next character start at once. 

- For rest after loop, there is diagonal line making an angle of 225 ± 20 

Shape at middle position

- Shape same as starting position joined with previous character . All conditions of shape at start location remain same.

Shape at end position

- Loop remains same as of middle position. After loop, shape is same as basic shape after loop. 

Wao class

One character is present named —Wao 

Basic Shape

First, there is loop same as —Fay and after loop there is a curve same as —Daall facing towards left.

Shape at start position

- Remain same as of isolated.

Shape at middle position

- Never comes in middle position of ligature

Shape at end position

- Remain same as of isolated

Gol Hay class

One character is present in this class named —Gol Hay

Basic Shape

- Its shape is same as circle,




starting from right moving left and



downward then right and upward back to starting position making a loop of radius 4 to 5 pixel.

- Alternatively, a curve of angle 235 ± 20 then moves right 2 to 3 pixel and at the end curve of angle 60 ± 20 . There is an intersection point making a loop.

Shape at start position


At start, its shape is same as shape of —Kashti at start position 

Shape at middle position


A line of angle 235 ± 20 followed by a curve of angle 60 ± 20 . There is cusp point between these two curves.

Shape at end position

At the end of ligature, its shape is a small line of angle 235 ± 20 followed by a small curve and a medium line of angle 150 ± 20 . When height of this line approaches starting point then again a curve followed by small line of angle

235 ± 20 

Do Chashmi Hay class

One character is present in this class named —Do chashmi Hay 

Basic shape

There is a small line of angle 235 ± 20 followed by a loop same as loop of —Swad. After first loop, there is another loop same as —Swad intersecting previous one and moving back to starting position. At the end, there is small horizontal line.

Shape at start position

Same as isolated shape


Shape at middle position

Same as isolated shape


Shape at end position

Same as isolated shape



ChotiYaa class

One character is present in this class named —ChotiYaa 

Basic Shape

A small line/curve of angle 245 ± 20 followed by a curve of angle 315 ± 20 . Then there is semicircle same as —Fay class,

Shape at start position

- When it comes in starting position its shape remain same as shape of —Kashti at starting position .
- All conditions of shape of —Kashti at starting position remain same here

Shape at middle position

- When it comes in starting position its shape remain same as shape of —Kashtil at middle position. کیا
- All conditions of shape of —Kashti at middle position remain same here

Shape at end position

- If ChotiYaa follows Kashti, —Jeem, —Fay, —Kaaf or —Laam class then its shape is only last part i.e. semicircle فی
- For rest its shape remains same as of isolated. صی

Bari Yaa class

One character is present in this class named —Bari Yaa ے

Basic Shape

A small line of angle 245 ± 20 followed by big horizontal line from left to right making an angle of zero.

- Or at start there is small vertical line from top to bottom and rest part is same as previous ے

Shape at start position

Same as shape of —ChotiYaa at starting position.

یا

Shape at middle position

Same as shape of —ChotiYaa at middle position.

کیا






Shape at end position

Its shape remains same as isolated

ے

Results and conclusion

We have tested the online Urdu OCR developed earlier by (Husain et al., 2007) by using 10 native Urdu writers and found out that there were a number of ambiguities in variations which result in failure in recognition by the engine. Moreover, the standard writing rules described above have not been devised keeping in view the online input and as such are not efficient in writing using a stylus or digital pen. It is therefore recommended that writing rules be modified keeping in view stroked based convenience and efficiency for online input devices.

Urdu language is difficult to learn and write. If proper writing rules are not followed in online input then recognition rate decreases as shape from  to  seems same but stroke sequence is completely changed. We also analyzed that there are many pen ups required in writing different ligatures which make recognition process slow, e.g. if we want to write , first we write  then a pen up and write the stroke. However, if we write this stroke without pen up then this will be more convenient and efficient i.e. .

Future direction of research

Future work includes devising efficient and convenient rules for online input. This will make recognition engine more efficient. Also, secondary strokes i.e., diacritics are not discussed in this paper, so a future direction of research would be to devise rules including diacritical marks.

References

- Aamir, A., Ayesha. A., Irfan, S., Sara and Sheraz. 2001. Contextual Shape Analysis of Nastalique. CRULP Annual Report.
- Sattar, A. 1985. Retrieved from http://eprints.ecs.soton.ac.uk/16510/1/ASattar_85.doc
- Mukhtar, O., Setlur and Srirangaraj. 2009. Experiments on Urdu Text Recognition. Guide to OCR for Indic. Scripts, Advances in Pattern Recognition, ISBN 978-1-84800-329-3. Springer-Verlag London. 163
- Reza, S. B., and Peyman, A. 2005. Nastaaliq handwritten word recognition using a continuous-density variable-duration Hmm. The Arabian Journal for Science and Engineering 30 (1B).
- Husain, S. A., Sajjad, A. and Anwar, F. 2007. Online Urdu Character Recognition Engine MVA 2007 IAPR Conference on Machine.
- Zaheer, A., Jehanzeb, K. O., Inam, S. and Awais A. December 2007. Urdu Nastaleeq Optical Character Recognition. In proceedings of world academy of science, engineering and technology 26.